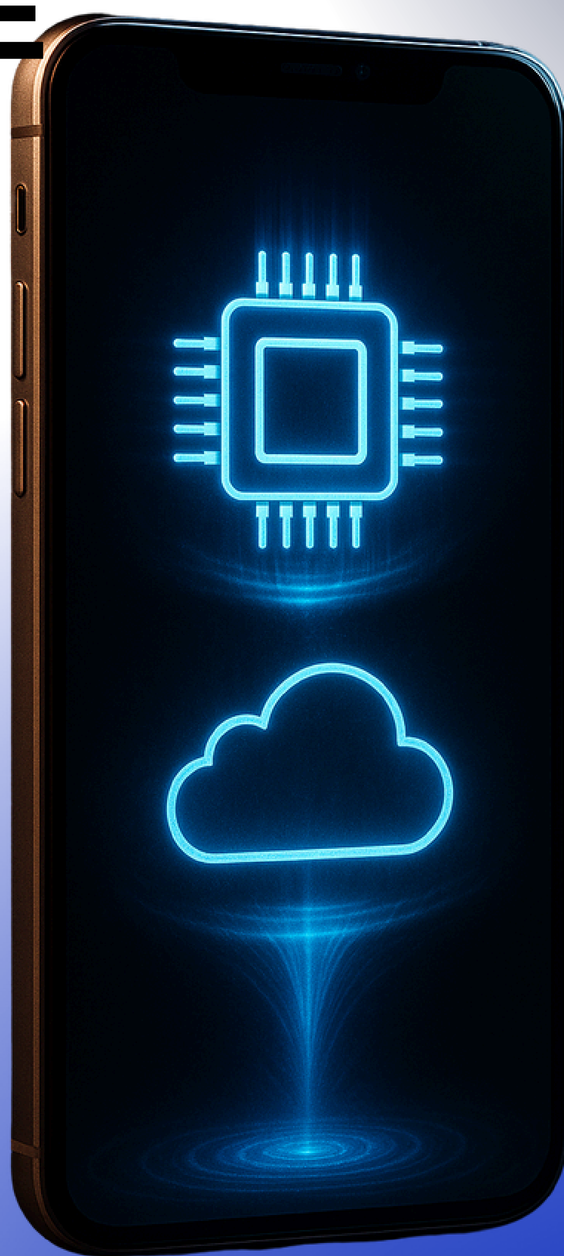
GUÍA DEFINITIVA: CÓMO USAR MODELOS DE LENGUAJE PEQUEÑOS (SLM) EN 2025

APRENDA A POTENCIAR SU DISPOSITIVO CON IA EFICIENTE, PRIVADA Y SIN INTERNET





Soporte360 IT

SOPORTE TECNICO

Abril 2025

¿Qué es SLM?

Los **modelos de lenguaje pequeños y abiertos** (SLM) son una revolución en 2025. A diferencia de los **LLM,** estos sistemas **pueden ejecutarse en dispositivos móviles** sin necesidad de conectarse a servidores masivos. Esto los hace **más accesibles y eficientes,** con aplicaciones en educación, traducción automática y diagnóstico médico en zonas con recursos limitados.

Parámetro	SLM (Local)	IA en la Nube
Tiempo respuesta	0.2 seg	1.5 seg
Privacidad	Datos 100% locales	Riesgo de filtraciones
Costo	Gratis/único pago	Suscripción mensual

Ejemplos de pequeños modelos lingüísticos

Los **profesionales utilizan los SLM** en muchas industrias porque **son ligeros, rápidos y no necesitan muchos recursos para funcionar.** Aquí se muestra a **algunos de estos modelos** con sus parámetros y características clave:

Nombre del modelo	Parámetros	Código abierto	Características principales
Qwen2	0,5B, 1B, 7B	Sí	Escalable, adecuado para diversas tareas
Mistral Nemo 12B	12B	Sí	Tareas complejas de PNL, despliegue local
Llama 3.1 8B	8B	Sí*.	Potencia y eficiencia equilibradas
Pythia	160M - 2,8B	Sí	Centrado en el razonamiento y la codificación
Cerebras-GPT	111M - 2,7B	Sí	Eficiente desde el punto de vista informático, sigue las leyes de escalado de Chinchilla
Phi-3,5	3.8B	Sí**	Contexto de gran longitud (128K tokens), multilingüe
StableLM-zephyr	3B	Sí	Inferencia rápida, eficiente para sistemas de borde
TinyLlama	1.1B	Sí	Eficaz para dispositivos móviles y de borde
MobileLLaMA	1.4B	Sí	Optimizado para dispositivos móviles y de bajo consumo
LaMini-GPT	774M - 1,5B	Sí	Multilingüe, tareas de seguimiento de instrucciones
Gemma2	9B, 27B	Sí	Despliegue local, aplicaciones en tiempo real
MiniCPM	1B - 4B	Sí	Rendimiento equilibrado, optimizado para inglés y chino
OpenELM	270M - 3B	Sí	Multitarea, baja latencia, eficiencia energética
DCLM	1B	Sí	Razonamiento de sentido común, deducción lógica
Fox	1.6B	Sí	Velocidad optimizada para aplicaciones móviles

*Con restricciones de uso
**Sólo con fines de investigación

No solo son chatbots

Pensemos en los **asistentes móviles,** esos **asistentes de voz** de su teléfono que lo ayudan a desenvolverse a lo largo del día. Los **SLM** lo hacen posible. Permiten la **predicción de texto en tiempo real,** los **comandos de voz** e incluso la traducción sin necesidad de enviar datos a la nube. Todo se hace **localmente,** lo que significa respuestas más rápidas e interacciones que **preservan más la privacidad.**

Por ejemplo**, SwiftKey y Gboard** utilizan **SLM** para proporcionar **sugerencias de texto** contextualmente precisas**,** lo que mejora la velocidad y la precisión de la escritura.

Esto **también se extiende a las aplicaciones offline,** en las que la IA puede seguir funcionando sin conexión a Internet, lo que la hace útil en zonas con conectividad limitada.

Google Translate, por ejemplo, ofrece funciones de **traducción offline** basadas en **SLM,** lo que facilita la comunicación en zonas con **acceso limitado a Internet.**



Gran consumo de los LLM (como ChatGPT)



Los **modelos de lenguaje de gran tamaño,** también conocidos como **LLM,** son modelos de aprendizaje profundo muy grandes que se preentrenan con grandes cantidades de datos. El transformador subyacente es un conjunto de redes neuronales que consta de un codificador y un decodificador con capacidades de autoatención. El codificador y el decodificador extraen significados de una secuencia de texto y comprenden las relaciones entre las palabras y las frases que contiene.

LLM más grandes, con más parámetros, consumen más energía tanto en el entrenamiento como en inferencia. Un aumento de 10 veces en el tamaño del modelo puede duplicar su consumo energético. **La generación de imágenes** consume hasta 2.907 kWh por cada 1.000 inferencias, mientras que **la generación de texto** es mucho más eficiente: 0,047 kWh por cada 1.000 respuestas.

¿Cómo ejecutar modelos SLM en el teléfono o computadora?

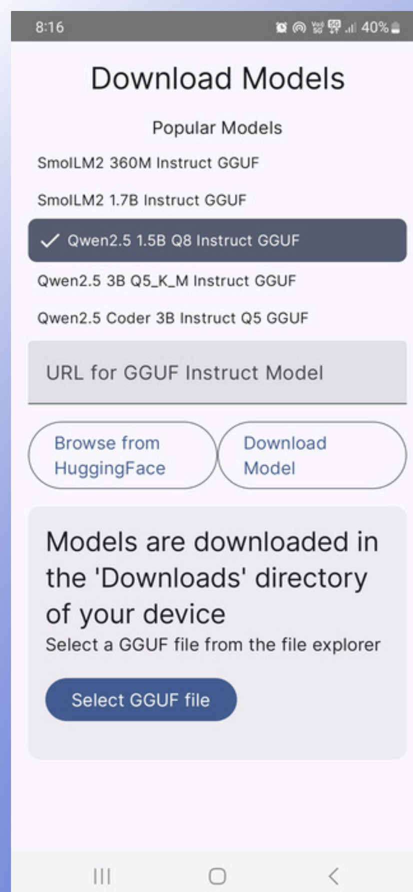
Hay aplicaciones tanto para las computadoras como para los teléfonos móviles que permiten correr **modelos SLM de manera local.**
Algunas de ellas son **LM Studio** para las computadoras en entorno **Windows, Linux y MacOS.**
Y para teléfonos móviles **Android** hay aplicaciones como **SmolChat.**

Tenga en cuenta que si bien son modelos pequeños, necesita un teléfono con capacidad para poder correrlo igualmente y con una versión moderna de Android.

El repositorio oficial es: https://github.com/shubham0204/SmolChat-Android
Puede **descargar la aplicación** desde: https://github.com/shubham0204/SmolChat-Android/releases donde se baja el **.apk**

Una vez instalada la aplicación SmolChat

Recuerde permitir instalaciones de **“fuentes desconocidas”** lo busca así en **ajustes** de su teléfono. Es para lograr instalar el **.apk.** Ya que es por fuera de la Play Store.
Luego de instalarla dentro de la aplicación podemos **descargar modelos,** que se almacenarán en la carpeta “Descargas”, de nuestro dispositivo.



Fuentes:

- https://blogs.unini.edu.mx/tecnologias-tics/2025/02/27/el-futuro-de-la-inteligencia-artificial-en-2025-avances-y-retos
- https://cobusgreyling.medium.com/run-a-small-language-model-slm-local-offline-1f62a6cbdaef
- https://github.com/shubham0204/SmolChat-Android
- https://www.datacamp.com/es/blog/small-language-models
- https://www.pressreader.com/
- https://aws.amazon.com/es/what-is/large-language-model/
- https://es.linkedin.com/pulse/pero-cu%C3%A1nto-gasta-realmente-un-llm-rafael-lopez-tl90f



Soporte360 IT

SOPORTE TECNICO

@SOPORTE360IT

WWW.SOPORTE360IT.COM